DOCUMENT RESUME

ED 303 083 HE 022 103

AUTHOR Menges, Robert J.
TITLE Why Research May Underestimate Effects of Feedback from Student Ratings. ASHE 1988 Annual Meeting Paper.
PUB DATE Nov 88
NOTE 10p.; Paper presented at the Annual Meeting of the Association for the Study of Higher Education (St. Louis, MO, November 3-6, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Instruction; Educational Research; *Feedback; Higher Education; Instructional Improvement; Performance; *Student Evaluation of Teacher Performance; Student Reaction; Teacher Evaluation; *Teacher Improvement; Teacher Response
IDENTIFIERS *ASHE Annual Meeting

ABSTRACT

Students on many campuses regularly evaluate their courses, and reports of these evaluations are sent to their instructors. Such feedback should result in improved teaching and in more positive evaluations, yet neither informal observation nor systematic studies reveal sizable feedback effects. The true effects of feedback of this kind are modest, and estimates of the true effects are attenuated by design and measurement problems and definition and implementation problems. Three kinds of evidence are described that support this contention: (1) original sin (some faculty deny the value of student ratings and once they have a PhD or get tenure, they no longer are able to distinguish pedagogical truth from pedagogical error); (2) archival data (little evidence of improved ratings over time which would indicate improvement); and (3) controlled studies, design and measurement, generalizing from midterm pretests, measurement scale characteristics, and definition and implementation problems, including definition of feedback, and unstated assumptions). The influence of generalizing from midterm pretests is uncertain. All of the other problems are likely to reduce the chances of identifying feedback effects. Measurement issues make it harder to detect true differences. Diffusion treatment can dilute feedback effects. Failure to verify implementation of treatment and investigate other assumptions about the treatment also attenuate true effects of feedback. Contains 8 references. (SM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

HE 022 103

ED303083

Why Research May Underestimate Effects of Feedback from Student Ratings[1]

Robert J. Menges

Northwestern University

2003 Sheridan Road

Evanston, IL 60208-2610

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ASHE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Abstract

Students on many campuses regularly evaluate their courses, and reports of these evaluations are sent to their instructors. Such feedback should result in improved teaching and in more positive evaluations, yet neither informal observation nor systematic studies reveal sizeable feedback effects. I believe the true effects of feedback of this kind are modest, but I also believe that estimates of the true effects are attenuated by two types of problems: design and measurement problems and definition and implementation problems.

[1]Paper presented at the meeting of the Association for the Study of Higher Education, St. Louis, November, 1988.

HE022103

ASHE ASSOCIATION FOR THE STUDY OF HIGHER EDUCATION

Texas A&M University
Department of Educational Administration
College Station, TX 77843
(409) 845-0393

This paper was presented at the annual meeting of the Association for the Study of Higher Education held at the Adam's Mark Hotel in St. Louis, Missouri, November 3-6, 1988. This paper was reviewed by ASHE and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC collection of ASHE conference papers.

Why Research May Underestimate Effects of Feedback from Student Ratings

The title of the symposium for which this paper was prepared asks why feedback from student ratings is not more effective for improving teaching. There are reasons for putting the question this way, rather than phrasing it, for example, "Is feedback from student ratings effective for improving teaching?" or "How effective is feedback from student ratings for improving teaching?" In this paper, I first discuss my reasons for contending that student ratings feedback has had only a modest influence toward improving teaching. Second, I argue that estimates of feedback effects are even lower than warranted because of problems with studies on this topic.

Evidence for the Modest Effect

Three kinds of evidence bear on my contention that there is only a modest effect of feedback from student ratings.

1. Original sin. With tongue in cheek, I term the first reason original sin. A number of people with whom I have discussed this topic express no surprise that routinely reporting end-of-term evaluations to faculty produces little improvement. I hear comments such as "Faculty don't believe they can learn anything from those questionnaires," or "Professors have a marvelous capacity to deny the truth of what students say," or "No matter what you tell them, faculty just teach the way they were taught." These comments suggest a view of faculty as persons who, once having tasted the fruits of Ph.D. knowledge or (even worse) once having attained tenure, enter a state of sin, and are no longer able to distinguish pedagogical truth from pedagogical error.

My own view is quite different. I believe that most of us are serious about our teaching, that we actively seek feedback of various kinds, and that we attempt to modify our teaching in light of new information. Of course, some faculty in some circumstances deny the value of student ratings, but I do not believe this condition to be inherent or universal. To be sure, it is not an adequate explanation for the modest effects of student ratings feedback.

2. Archival data. For some years on many campuses there have been end-of-term evaluations by students followed by feedback to instructors. If feedback about one's work leads to modifying and improving teaching, archival data at these institutions should show gradually escalating ratings. This does not seem to be happening. There are individual cases where improvement occurs and sometimes that improvement is dramatic, but I have been unable to find data documenting widespread and strong increases in ratings over time.

My impression that only small improvement occurs over time is shared by people responsible for well known student ratings systems. Bill Cashin, for example, tells me that he has noticed an increase in average ratings on the IDEA form when faculty are evaluated in the same or similar courses, but that the increase is of little practical significance. If his clients say they want to use the IDEA system to improve teaching, he advises them to offer consultation as part of their program (personal communication, September 28, 1988).

Two published studies include reports of archival data. Hoyt and Howard (1978) used as a no-treatment control group 263 faculty at Kansas State who had been evaluated for the same course on at least two occasions over three or four years. They received ratings feedback but no other instructional improvement services. Their later ratings showed a gain that "while statistically significant, was not dramatic in the absolute sense. On most measures, the average participant improved about one-fifth of a standard deviation" (p. 30).

In a study at the University of Illinois, Urbana-Champaign, Gray and Brandenburg (1985) tracked evaluations with feedback of 304 instructors over four semesters. They found only "minor improvement over time" regardless of whether teachers used the standard evaluation form or created their own form by choosing from a catalogue of available items. Most of the small improvement occurred between the first and second evaluations. Since many of these teachers were receiving student ratings feedback for the first time (at least with the system used in the study), I infer that feedback may have its greatest impact the first time it is experienced.

3. Controlled studies. There have been some thirty controlled studies on this topic. These are studies in which summaries of student ratings are provided to some faculty but not to others and subsequent ratings of these groups are compared. Meta-analyses of this literature (Cohen, 1980; Menges & Brinko, 1986) agree that effects of these statistical summaries of student ratings are modest. The average instructor not receiving feedback is, of course, at the 50th percentile. The average instructor who receives feedback from ratings is at about the 59th percentile, an improvement that some may argue barely justifies costs of processing rating forms.

According to these meta-analyses, feedback effects are considerably enhanced when ratings are accompanied by consultation, that is, when faculty discuss their evaluations with someone else. Although the number of studies with a consultation condition is not large, their findings are in substantial agreement. The average faculty member who receives both student ratings feedback and consultation is subsequently rated at about the 85th percentile. Our recent replication of the meta-analysis with a slightly different group of studies reaches the same conclusion (L'Hommedieu, Menges, & Brinko, 1988a). The impressive size of this consultation effect has stimulated research on consultative interactions (Brinko, 1989).

Problems with Controlled Studies

In this paper I am concerned with studies of feedback solely from student ratings, apart from consultation or other interventions. I believe that the effects of written statistical summaries, though modest, may be greater than studies reveal. Some of the problems with these studies lie with measurement and design issues and others with how the construct of feedback is defined and implemented. (Fuller discussion of these and other problems is in L'Hommedieu, Menges, & Brinko, 1988b.)

Design and Measurement Problems

In the first set of problems are issues of research design and instrumentation.

1. Generalizing from midterm pretests. The typical design is a two-group pretest/posttest comparison. Experimental group teachers receive feedback from an evaluation

administered at midterm and are again rated at the end of the term. Their end-of-term ratings are compared with those of control group teachers who were also rated at midterm but who did not receive feedback. This design is vulnerable to pretest effects, although such effects have been ruled out in several studies that include no-pretest comparison groups. More serious is the external validity problem of generalizing from this research to campus programs that, rather than using midterm ratings with immediate feedback, employ end-of-term ratings with feedback that occurs after the course being rated has concluded. How does the salience of feedback differ in these situations? How different are these situations in the kind and amount of information available to students for evaluating the course and in the levels of stress under which teachers and students are operating? Although this midterm pretest design offers important internal validity advantages, it compromises external valitidy in ways impossible to estimate; further, it threatens construct validity.

2. Measurement scale characteristics. Most instruments used in these studies employ the Likert Scale. Midterm ratings are nearly always above the midpoint of the scale, thus limiting room for future improvement. Aggregating Likert Scale data across groups, times, situations, and even across studies (as we did in the meta-analyses) compromises assumptions required by the scale. An additional measurement problem is researchers' failure to report reliability data when instruments are constructed locally for a particular study.

3. Diffusion of treatment. As suggested earlier, it seems reasonable that feedback effects would be strongest the first few times one teaches a particular course or works with particular kinds of students. Researchers should take care to note how much previous experience participating teachers have had with student ratings feedback. In at least a third of these studies, feedback systems were already institutionalized; in other studies the issue was not addressed. Thus in an unknown proportion of these studies, experimental group teachers may be past the period of optimal benefit and teachers in control groups may already have benefited from feedback. Failure to isolate the experimental treatment constitutes a grave threat to the validity of these studies.

Definition and Implementation Problems

The final set of problems has to do with how *feedback*, the construct of interest in these studies, is defined and implemented.

1. Definition of feedback. In order to design a feedback treatment, researchers should explicitly state the necessary and sufficient conditions a message must possess if it is to constitute *feedback* and if it is to produce the intended effect. From writings on improving teaching (for example, McKeachie, 1986, and Fuhrmann & Grasha, 1983) desirable characteristics of feedback messages can be postulated. Feedback should provide information that is new to the teacher; the information should be so specific that correctives can be derived from it; and the feedback message should arouse only moderate anxiety, that is, the message should be strong enough to challenge but not so strong as to threaten one's self-concept. Rarely do researchers define the feedback construct this clearly. Nor are feedback messages described in sufficient detail to determine what criteria they were designed to meet.

2. Implementation of treatment. To experience the treatment as intended, instructors in feedback groups must receive summaries of student evaluations in a timely fashion and must give some attention to them. Unfortunately, these studies seldom include data verifying that feedback summaries have been received and attended to. By contrast, consultation, when it occurs, assures that some attention has been paid and that some information has been processed. Indeed, this may account in large measure for the positive effects of consultation.

3. Unstated assumptions. Beyond the assumption that experimental group members have received and attended to feedback messages, other assumptions are apparently held by researchers. Here are three examples of such assumptions. (a) That teachers receiving feedback have a common understanding of the terms used on evaluation forms. These terms include *explains*, *discusses*, *grades fairly*, *agree*, *strongly agree*, and so on. (b) That feedback recipients share the same understanding of numbers used in reports, of distances between those numbers, and of such terms as *mean* and *standard deviation*. (c) That all instructors in the experimental

groups want to have high ratings. It is important to demonstrate the truth of these assumptions empirically before claiming that it is *feedback* that explains differences obtained between experimental and control groups.

Conclusion

My contention is that these problems with studies of feedback from student ratings mask the true effects of feedback. To conclude, I review each of the problems with regard to the direction in which it is likely to affect research results.

Generalizing from midterm pretests could be argued either as inflating or deflating results, so I call the direction of its influence uncertain. All of the other problems, however, are likely to reduce chances of identifying feedback effects. Measurement issues, such as ceiling effect and restriction of range make it more difficult to detect true differences. Diffusion of treatment is almost certain to dilute feedback effects. Lack of explicitness in defining *feedback* likely minimizes differences and makes research difficult to interpret and to replicate. Failure to verify implementation of treatment and to investigate other assumptions about the treatment likewise attenuate true effects of feedback.

I do not mean to imply that these researchers have been careless or uninformed or that their efforts have been wasted. There is some admirable work here, often done under difficult practical circumstances. But these studies do not reveal as much as we would like to know about the value of student ratings feedback for improving teaching.

References

Brinko, Kathleen T. (1989). Instructional consultation with feedback in higher education. Journal of Higher Education, accepted for publication.

Fuhrmann, B. S., & Grasha, A. F. (1983). A practical handbook for college teachers. Boston: Little, Brown.

Gray, D. M., & Brandenburg, D. C. (1985). Following student ratings over time with a catalog-based system. Research in Higher Education, 22, 155-168.

Hoyt, D. T., & Howard, G. S. (1978). The evaluation of faculty development programs. Research in Higher Education, 8, 25-38.

L'Hommedieu, R., Menges, R. J., & Brinko, K. T. (1988a). The effects of student ratings feedback to college teachers: A meta-analysis and review of research. Evanston, IL: Center for the Teaching Professions, Northwestern University.

L'Hommedieu, R., Menges, R. J., & Brinko, K. T. (1988b). Methodological explanations for the modest effects of student ratings feedback. Paper presented at the American Educational Research Association, New Orleans.

McKeachie, W. J. (1986). Teaching tips. Lexington, MA: D. C. Heath.

Menges, R. J., & Brinko, K. T. (1986). Effects of student evaluation feedback: A meta-analysis of higher education literature. Paper presented at the American Educational Research Association, San Francisco.